Causal inference and machine learning approaches for evaluation of the health impacts of large-scale air quality

Evaluation of the health impacts of the 1990 Clean Air Act Amendments using causal inference and machine learning

Rachel C. Nethery[1], Fabrizia Mealli[2], Jason D. Sacks[3], Francesca Dominici[1] *

Abstract

We develop a causal inference approach to estimate the number of adverse health events that were prevented due to changes in exposure to multiple pollutants attributable to a large-scale air quality intervention/regulation, with a focus on the 1990 Clean Air Act Amendments (CAAA). We introduce a causal estimand called the Total Events Avoided (TEA) by the regulation, defined as the difference in the number of health events expected under the no-regulation pollution exposures and the number observed with-regulation. We propose matching and machine learning methods that leverage population-level pollution and health data to estimate the TEA. Our approach improves upon traditional methods for regulation health impact analyses by formalizing causal identifying assumptions, utilizing population-level data, minimizing parametric assumptions, and collectively analyzing multiple pollutants. To reduce model-dependence, our approach estimates cumulative health impacts in the subset of regions with projected no-regulation features lying within the support of the observed with-regulation data, thereby providing a conservative but data-driven assessment to complement traditional parametric approaches. We analyze the health impacts of the CAAA in the US Medicare population in the year 2000, and our estimates suggest that large numbers of cardiovascular and dementia-related hospitalizations were avoided due to CAAA-attributable changes in pollution exposure.

Keywords: Matching, Bayesian Additive Regression Trees, Counterfactual Pollution Exposures, 1990 Clean Air Act Amendments

*1=Department of Biostatistics, Harvard T.H. Chan School of Public Health, 2=Department of Statistics, Computer Science, Applications, University of Florence, 3=National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency. The authors gratefully acknowledge funding from NIH grants 5T32ES007142-35, R01ES024332, R01ES026217, P50MD010428, DP2MD012722, R01ES028033, R01GM111339, R01HD092580, R01AG060232-01A1 and R01MD012769; HEI grant 4953-RFA14-3/16-4; and EPA grant 83587201-0.

1 Introduction

In its 2011 cost-benefit analysis of the 1990 Clean Air Act Amendments (CAAA), the United States Environmental Protection Agency (EPA) estimates that the total direct costs of compliance in the year 2000 were nearly $20 billion, and it anticipates that these costs will increase to over $65 billion by 2020 (US EPA, 2011). The CAAA is an expansion of the 1970 Clean Air Act, and it has prompted the enactment of numerous new regulatory programs, both at national and local levels, to ensure that pollution emissions limits are observed and that air quality standards are being met (see Henneman et al. (2019) for a summary). Hereafter, for simplicity we use the term CAAA to refer to the set of regulations put in place to adhere to the law. While the EPA estimates that the economic benefits of the CAAA dwarf the costs (estimated 2000 benefits $770 billion, estimated 2020 benefits $2 trillion), the increasing compliance costs call for continued evaluations of the effects of the CAAA using the most advanced, rigorous methods. Of particular interest to the public are the health impacts of the CAAA.

Traditionally, assessments of the health impacts of large-scale regulations, including assessments of the CAAA, have relied on the combination of simulated air quality models and pollutant-health exposure-response functions (ERF). These assessments have relied on various tools including the World Health Organization's AirQ+ software (WHO, 2019) and EPA's Environmental Benefits Mapping and Analysis Program - Community Edition (BenMAP-CE) software (US EPA, 2011; Sacks et al., 2018). For example, when examining a regulation within BenMAP-CE, atmospheric chemistry modeling is first used to estimate concentrations of a pollutant on a grid across the area of interest in a post-regulation year under (1) the factual/observed scenario of regulation implementation and (2) the counterfactual/unobserved scenario of no regulation (hereafter referred to as factual and

counterfactual exposures). The difference in the factual and counterfactual pollutant levels in each grid cell are used as inputs in the ERFs. The ERFs employ these differences along with the size of the exposed population, the baseline rate of mortality or morbidity, and a health effect estimate, i.e., a linear model coefficient from a previously published epidemiologic study capturing the relationship between pollutant exposure and a health outcome, to estimate the number of health events prevented by the regulation-attributable changes in pollutant exposures in the specified year. This process is performed separately for each relevant pollutant-health outcome combination. For more detail on the traditional approach and the ERFs, see Section 1 of the Supplementary Materials.

Causal inference principles have historically been featured heavily in analyses of the health effects of short-term air pollution interventions, such as traffic bans often imposed in cities hosting Olympic games, which can often be formulated as natural experiments. Larger, gradually-implemented regulatory actions, such as the CAAA, are more complex because 1) the resultant changes in levels of multiple pollutants vary in space and time, 2) long term health trends may coincide with changes in air quality, and 3) time-varying confounding is likely (van Erp et al., 2008). Only recently have causal inference approaches begun to be developed to address these issues. Zigler et al. (2012) investigate the effect of a component of the CAAA, the National Ambient Air Quality Standard (NAAQS) non-attainment designations, on health in the Medicare population using a principal stratification approach. To our knowledge, factual and counterfactual exposures from air quality modelling software and causal inference methodology have not yet been integrated for regulation evaluation, despite their natural connection.

In this paper, we seek to estimate the number of health events prevented by the CAAA in a specific year using a unique approach that combines factual and counterfactual pollution

exposures with observed health outcome data from the Medicare population (instead of relying on ERFs derived from previously conducted epidemiologic studies). We intend to answer the question "How many mortality events, cardiovascular hospitalizations, and dementia-related hospitalizations were prevented in the Medicare population in the year 2000 due to CAAA-attributable changes in particulate matter ($PM_{2.5}$) and ozone (O_3) exposures in the same year?" Because pollution exposures may continue to impact health beyond the year of exposure, we also wish to determine how many of each of these health events were prevented in the year 2001 due to the CAAA-attributable changes in pollutant exposures in the year 2000. We rely on estimates of $PM_{2.5}$ and O_3 exposure levels across the US for the year 2000 under the factual with-CAAA scenario and under the counterfactual no-CAAA scenario. We combine them with zipcode level counts of mortality, cardiovascular hospitalizations, and dementia-related hospitalizations from Medicare in the years 2000 and 2001. Then we utilize the number of health events observed under factual $PM_{2.5}$ and O_3 levels to inform estimation of the number of health events that would have occurred under the counterfactual pollutant levels (the counterfactual outcome).

We introduce a causal inference framework that can be applied to evaluate the CAAA or any other large-scale air quality regulation. Reliance on counterfactual pollution predictions, which come from the EPA's cost-benefit analysis of the CAAA (US EPA, 2011), enables us to conduct a causal inference investigation in this otherwise intractable setting where we observe no data whatsoever under the no-regulation scenario. The first novel feature of our work is the introduction of a causal estimand, which we call the Total Events Avoided (TEA) by the regulation. It is defined as the sum across all units of the difference in the expected number of health events under the counterfactual pollution exposures and the observed number of health events under the factual pollution exposures. We also lay

out the corresponding identifiability conditions. The second novel aspect of this paper is in the development of a matching approach for estimation of the TEA. We also propose an alternative estimation approach using machine learning methods. Relying on minimal modeling assumptions, both of these approaches use confounder-adjusted relationships between observed pollution exposures and health outcomes to inform estimation of the counterfactual outcomes. While we are seeking to estimate the same quantity estimated in traditional health impact analyses of regulations (the number of health events prevented in a year due to regulation-attributable changes in pollutant exposures that year), the statistical methods used are quite different. Our approach improves on the traditional one by: 1) defining the causal parameter of interest; 2) formalizing the assumptions needed to identify it from data; 3) relying on population-level health outcome data for estimation; 4) minimizing parametric assumptions; and 5) accounting for synergistic effects of multiple pollutants. However, to avoid extrapolation and heavy reliance on parametric modeling assumptions, our methods exclude some areas from the analysis, thereby producing conservative yet data-driven estimates of the health impacts of the regulation.

Matching is one of the most commonly used approaches to estimate causal effects (Ho et al., 2007; Stuart, 2010). Machine learning procedures have emerged more recently as a tool for causal inference (Hill, 2011; Hahn et al., 2020). Both have primarily been used to estimate average treatment effects in settings with a binary treatment, with recent limited extensions to the continuous exposure setting (Kreif et al., 2015; Wu et al., 2018). To our knowledge, neither approach has been used in the context of a multivariate continuous treatment (the pollution exposures in our setting). To estimate the TEA, we introduce both a matching method and an adaptation of the Bayesian Additive Regression Trees (BART) algorithm (Chipman et al., 2010) that accommodate multivariate continuous treatments.

In Section 2, we discuss the air pollution and Medicare data that motivate our methodological developments. In Section 3, we formally introduce the TEA and identifying assumptions, and we present our matching and machine learning methods for TEA estimation. Section 4 describes simulations conducted to evaluate the performance of these methods. In Section 5, we apply these methods to investigate the health impacts of the CAAA. Finally, we conclude with a summary and discussion of our findings in Section 6.

2 Data

In this paper, we focus on the number of health events prevented due to CAAA-attributable changes in two major pollutants– $PM_{2.5}$ and O_3. These two pollutants are known to have large health impacts. We have obtained state-of-the-art factual (with-CAAA) and counterfactual (no-CAAA) gridded $PM_{2.5}$ and O_3 exposure estimates for the continental US in the year 2000. $PM_{2.5}$ exposures (both factual and counterfactual) are measured in $\mu g/m^3$ and represent annual averages, while O_3 exposures are measured in parts per billion (ppb) and represent warm season averages. Our factual pollution exposure estimates come from so-called hybrid models which combine ground monitor data, satellite data, and chemical transport modeling to estimate exposures on a fine grid across the US. The factual $PM_{2.5}$ exposure estimates employed here are introduced in van Donkelaar et al. (2019) and are produced at approximately 1 km^2 grid resolution. Our factual O_3 exposure estimates were developed by Di et al. (2017), also at 1 km^2 grid resolution.

We employ the year-2000 counterfactual gridded $PM_{2.5}$ and O_3 exposure estimates for the continental US from the EPA's Second Section 812 Prospective Analysis (hereafter called the Section 812 Analysis), its most recent cost-benefit analysis of the CAAA (US

EPA, 2011). To produce these, gridded hourly emissions inventories were first created for the counterfactual (no-CAAA) scenario in 2000, representing estimated emissions with scope and stringency equivalent to 1990 levels but adjusted to economic and population changes in 2000. Note that this approach to creating counterfactual emissions inventories assumes that in the absence of the CAAA 1) no new emissions regulations would have been implemented in the US between 1990 and 2000 and 2) the scope and stringency of US emissions would not have increased. We anticipate that these assumptions would slightly bias the emissions estimates in opposite directions; thus, we expect that they balance out to create a realistic counterfactual emissions scenario. The emissions inventories were fed into atmospheric chemistry modeling software to produce estimates of counterfactual annual average $PM_{2.5}$ at 36 km^2 grid resolution, and counterfactual warm season O_3 at 12 km^2 grid resolution. For more detail on the creation of the counterfactual exposure estimates, see US EPA (2011). We note that factual exposure estimates were also produced for the EPA's Section 812 Analysis in an analogous manner to the counterfactual exposures. For justification of our choice to use factual exposures from the hybrid models instead, and potential impacts of this choice, see Section 2.1 of the Supplementary Materials.

Using the R package `gstat` (Gräler et al., 2016), we apply local block kriging to aggregate the gridded $PM_{2.5}$ and O_3 values to zipcodes, in order to merge them with the Medicare data (Cressie, 1993). See the zipcode maps of the factual and counterfactual $PM_{2.5}$ and O_3 data in Figure 1. In the analyses in the main manuscript, we make the strong assumption that the observed and counterfactual pollution levels are known or estimated without error, because current limitations in air quality modeling impede reliable quantification of uncertainties (Nethery and Dominici, 2019). This is a commonly-used assumption in air pollution epidemiology; however, we discuss the implications in Section 6 and explore

potential impacts using simulated exposure errors in the Supplementary Materials.

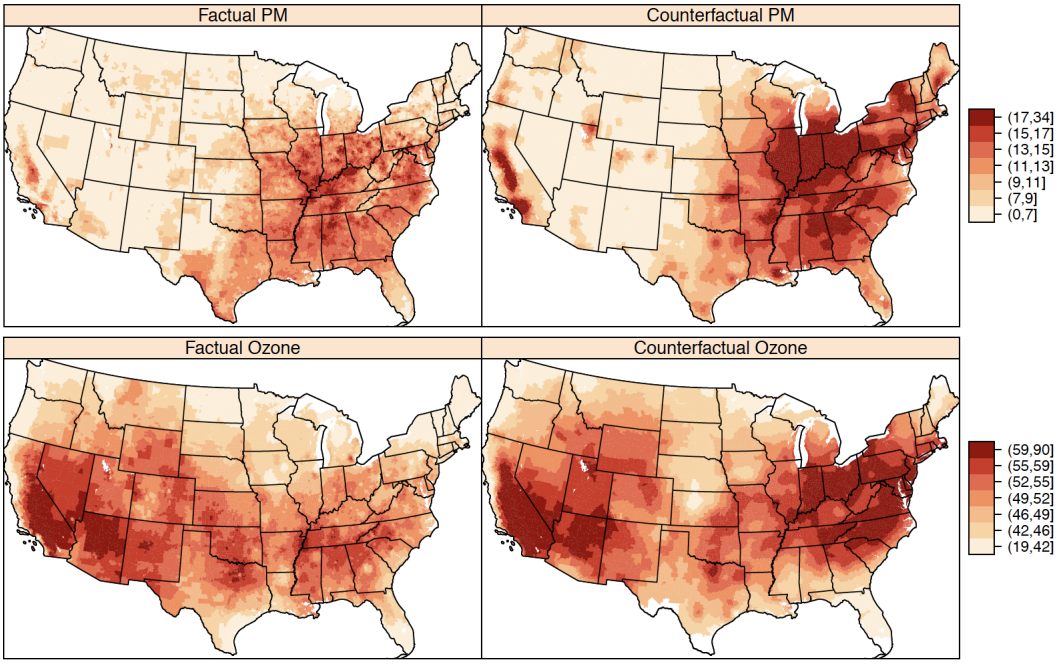

Figure 1: Maps of estimated year-2000 zipcode level factual and counterfactual annual average $PM_{2.5}$ ($\mu g/m^3$) and warm season average O_3 (ppb). The factual pollutant values are estimates of the true exposures with the CAAA. The counterfactual values reflect the anticipated pollutant exposures under the emissions scenario expected without the CAAA.

We construct datasets containing mortality records for all Medicare beneficiaries and hospitalization records for all Medicare fee-for-service (FFS) beneficiaries in the continental US for the years 2000 and 2001. Medicare covers more than 96% of Americans age 65 and older (Di et al., 2017). Because cross-zipcode moving rates are low in the Medicare population (Di et al., 2017; Abu Awad et al., 2019), we make the assumption that, for each of the 2000 and 2001 cohorts, subjects were exposed to the year-2000 pollution levels of their zipcode of residence. Zipcode counts of mortality, cardiovascular hospitalizations, and dementia-related hospitalizations in each cohort will serve as the outcomes in our analysis.

For detailed information about how these counts were constructed, including the ICD-9 codes used to classify hospitalizations, see Section 2.2 of the Supplementary Materials. We selected these health outcomes on the basis of previous literature associating each of them with air pollution (Pope III et al., 2002; Brook et al., 2010; Power et al., 2016). We also compute the Medicare person-months in each zipcode in each year to create rates of each health event (e.g. count of hospitalizations in the Medicare FFS population in the zipcode in year 2000/total Medicare FFS person-months in the zipcode in year 2000).

Finally, a set of potential confounders of air pollution and health relationships is constructed using data from the 2000 US census. All confounders are zipcode-aggregate features that reflect all residents of the zipcodes, not only Medicare beneficiaries. They are percent of the population below the poverty line (poverty), population density per square mile (popdensity), median value of owner-occupied properties (housevalue), percent of the population black (black), median household income (income), percent of housing units occupied by their owner (ownhome), percent of the population hispanic (hispanic), percent of the population with less than a high school education (education), and regional indicators (northeast, midwest, south, west). In Section 2.3 of the Supplementary Materials, we discuss our choice to conduct analyses at the zipcode level as opposed to coarser aggregations.

3 Methods

3.1 Estimand and Identifying Assumptions

In this section, we propose a causal inference estimand to measure the difference in the number of health events in a given year under the regulation and no regulation scenarios, and we lay out the identifying assumptions. The units of analysis for our case study of the

CAAA will be zipcodes, but throughout this section we use the term units to emphasize the generality of the approach to any areal units. Let Y_i denote the count of health events observed in unit i, $i = 1, ..., N$, and P_i the at-risk person-time in unit i. Then let $Y_i^* = \frac{Y_i}{P_i}$ be the event rate. While our estimand involves counts, we must conduct modeling with rates instead of counts when at-risk person-time varies. Let $\boldsymbol{X}_i$ denote a vector of observed confounders of the relationship between pollution and the health outcomes under study, as well as any factors besides the observed pollutants which were impacted by the regulation. Let $\boldsymbol{T}_i$ denote the Q-length vector of continuous pollutant exposure measurements for unit i, where Q is the number of pollutants under consideration (annual/warm season averages), and we assume that $\boldsymbol{T}$ has compact support over a Q-dimensional hyperrectangle, $\mathbb{Z}$, which is a subspace of $\mathbb{R}^Q$. $\boldsymbol{T}_i$, the vector of pollutant exposure levels, will serve as the "treatment" variable in our causal inference framework.

We develop our approach within the potential outcomes framework of Rubin (1974). Recall that we will make use of both the factual and counterfactual pollution levels for each unit. Then let $Y_i(\boldsymbol{T} = \boldsymbol{t}_{1i})$ be the number of health events that unit i would experience under its observed pollutant values ($\boldsymbol{t}_{1i}$) in the factual scenario of regulation implementation. Let $Y_i(\boldsymbol{T} = \boldsymbol{t}_{2i})$ be the number of health events that unit i would experience under its counterfactual pollution levels ($\boldsymbol{t}_{2i}$) in the no-regulation scenario, with all other features of the unit identical to the observed ones. For each unit, $Y_i(\boldsymbol{T} = \boldsymbol{t}_{1i})$ is observed and $Y_i(\boldsymbol{T} = \boldsymbol{t}_{2i})$ is unobserved. These potential outcomes can be constructed under the stable unit treatment value assumption (SUTVA) (Rubin, 1980). SUTVA requires that the health effect of a given treatment (pollution) level is the same no matter how that treatment level is arrived at, and that one unit's pollution level does not affect the number of health events in another unit (the latter is known as the no interference assumption). The former as-

sumption could be violated if, for instance, certain sources of pollution are more toxic than others. No interference is likely to be a strong assumption (Zigler et al., 2012; Zigler and Papadogeorgou, 2018), as most individuals are regularly exposed to the pollution levels in areas other than their area of residence, which could impact their health. More discussion of this assumption is provided in Sections 5 and 6.

Before formalizing the TEA, we emphasize that the TEA is designed to quantify the health impacts of the regulation only through resultant changes in the pollutants in $\boldsymbol{T}$, while holding all other measured features fixed at the observed levels (under the regulation). The TEA characterizes a counterfactual scenario in which the only changes from the observed with-regulation scenario are the pollution exposures, and compares the number of health events in this counterfactual scenario and the observed one. We explain the motive behind this estimand. Previous causal inference analyses (Zigler et al., 2012) have targeted the effects of air pollution regulations on health both via the resultant changes in pollutant exposures (associative effects) and via other intermediates (dissociative effects). These studies investigated an element of the CAAA, the NAAQS non-attainment designations, which impacted some locations and not others, providing variation in attainment status and observed health outcome data under both the attainment and non-attainment scenarios. These data enabled the estimation of both associative and dissociative effects. Here we instead wish to evaluate a larger regulation which had "universal" impacts in areas under its purview (hereafter referred to as a universal regulation). We do not observe any health outcome data whatsoever under the no-regulation scenario. With this limitation, our analysis must rely entirely on the observed outcome data under regulation. Because the only information we have about the no-regulation scenario is the counterfactual pollutant exposures, it is only through these pollutants that the observed health outcome data can inform

estimation of the health outcomes in the no-regulation scenario. Thus these data do not allow any investigation of dissociative effects, which is why we have defined the TEA as the effect of the regulation only through the resultant changes in the pollutants in $\boldsymbol{T}$. The Section 812 Analysis also estimates health effects exclusively through regulation-attributable changes in pollutants. Previous studies have found positive dissociative health impacts of regulations (Zigler et al., 2012). If such effects exist, the TEA will under-represent the total number of health events prevented by the regulation.

The number of health events prevented due to regulation-induced changes in pollutants is $\sum_{i=1}^{N} \{Y_i(\boldsymbol{T} = \boldsymbol{t}_{2i}) - Y_i(\boldsymbol{T} = \boldsymbol{t}_{1i})\}$. Because $Y_i(\boldsymbol{T} = \boldsymbol{t}_{2i})$ is unobserved for all i, we instead focus our analysis around the following estimand:

$$\tau = \sum_{i=1}^{N} \{\mathbb{E}(Y(\boldsymbol{T} = \boldsymbol{t}_{2i})|\boldsymbol{X}_i) - Y_i(\boldsymbol{T} = \boldsymbol{t}_{1i})\}$$

τ is the TEA. With $Y_i(\boldsymbol{T} = \boldsymbol{t}_{1i})$ observed for all i, we only need to estimate $\mathbb{E}(Y(\boldsymbol{T} = \boldsymbol{t}_{2i})|\boldsymbol{X}_i)$ for each i to obtain an estimate of the TEA. Note that this quantity is conditional on $\boldsymbol{X}_i$, formalizing our statement above that, in using the TEA to evaluate a regulation, we are making the assumption that the only factor that would have been different in the no-regulation scenario is the pollutant exposures. Moreover, this causal estimand is unique because it captures the health effects due to changes in multiple continuous pollutants simultaneously. This feature provides an important improvement over traditional regulation health impact analyses, where health impacts are estimated separately for each pollutant. When analyzing each pollutant separately, estimates can be biased if pollution exposures are correlated and may fail to capture synergistic effects. We provide a detailed comparison of the TEA to existing causal estimands in Section 3 of the Supplementary Materials.

We now present the assumptions needed to identify $\mathbb{E}(Y(\boldsymbol{T} = \boldsymbol{t}_{2i})|\boldsymbol{X}_i)$ from the observed data. These assumptions are extensions of those of Wu et al. (2018).

Assumption 1 (A1) Causal Consistency: $Y_i(\boldsymbol{T} = \boldsymbol{t}) = Y_i$ if $\boldsymbol{T}_i = \boldsymbol{t}$

The causal consistency assumption states that the observed outcome should correspond to the potential outcome under the observed treatment value.

Assumption 2 (A2) Weak Unconfoundedness: Let $I_i(\boldsymbol{t})$ be an indicator function taking value 1 if $\boldsymbol{T}_i = \boldsymbol{t}$ and 0 otherwise. Then weak unconfoundedness is the assumption that $I_i(\boldsymbol{t}) \perp\!\!\!\perp Y_i(\boldsymbol{T} = \boldsymbol{t})|\boldsymbol{X}_i$.

The weak unconfoundedness assumption was introduced by Imbens (2000) and is commonly used for causal inference with non-binary treatments. In our context, it says that assignment to treatment level $\boldsymbol{t}$ versus any other treatment level is independent of the potential outcome at treatment $\boldsymbol{t}$ conditional on the observed confounders.

Assumption 3 (A3) Overlap: For each $\{\boldsymbol{t}_{2i}, \boldsymbol{x}_i\}$, $0 < P(\boldsymbol{T} = \boldsymbol{t}_{2i}|\boldsymbol{X} = \boldsymbol{x}_i)$.

The overlap assumption in this context differs slightly from the overlap (or positivity) assumption in classic causal inference analyses. It says that, for each unit's set of counterfactual treatment and confounders, the probability of observing that treatment and confounder level together is greater than zero. This ensures that we are not considering counterfactuals outside the space of feasible treatment and confounder combinations.

Assumption 4 (A4) Conditional Smoothness: Let $\Theta_{\boldsymbol{t}} = [t_1 - \delta_1, t_1 + \delta_1] \times \cdots \times [t_Q - \delta_Q, t_Q + \delta_Q]$, t_q represents the qth element of $\boldsymbol{t}$, and $\delta_1, ..., \delta_Q$ are positive sequences tending to zero. Then

$$\lim_{\delta_1,...,\delta_Q \to 0} \mathbb{E}(Y|\boldsymbol{X}, \boldsymbol{T} \in \Theta_{\boldsymbol{t}}) = \mathbb{E}(Y|\boldsymbol{X}, \boldsymbol{T} = \boldsymbol{t})$$

This multidimensional smoothness assumption is needed due to the continuous nature of the multivariate treatments, which means that we will never have $\boldsymbol{T} = \boldsymbol{t}$ exactly and must instead rely on $\boldsymbol{T}$ within a small neighborhood of $\boldsymbol{t}$.

Using (*A1-4*), we show that $\mathbb{E}(Y(\boldsymbol{T} = \boldsymbol{t}_{2i})|\boldsymbol{X}_i)$ can be identified from observed data.

$$\begin{aligned}
\mathbb{E}(Y(\boldsymbol{T} = \boldsymbol{t}_{2i})|\boldsymbol{X}_i) &= \mathbb{E}(Y(\boldsymbol{T} = \boldsymbol{t}_{2i})|\boldsymbol{X}_i, \boldsymbol{T} = \boldsymbol{t}_{2i}) \quad \text{(by A2)} \\
&= \mathbb{E}(Y|\boldsymbol{X}_i, \boldsymbol{T} = \boldsymbol{t}_{2i}) \quad \text{(by A1)} \\
&= \lim_{\delta_1,...,\delta_Q \to 0} \mathbb{E}(Y|\boldsymbol{X}_i, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2i}}) \quad \text{(by A4)}
\end{aligned} \tag{1}$$

The expectation in the bottom line of equation 1 can be estimated from observed data for small, fixed values of $\delta_1, ..., \delta_Q$.

In the next two sections, we introduce methods to perform this estimation with the data described in Section 2. Before doing so, we clarify a few additional assumptions that will be relied upon in order to interpret a TEA estimate as the number of health events avoided due to the regulation-attributable changes in pollutants. These assumptions, which are not common in causal inference analyses, are needed in the universal regulation scenario due to the complete reliance on observed data under the regulation for estimation. First, we must assume that each of the following relationships would be the same in the regulation and no-regulation scenarios: (1) the pollutant-outcome relationships, (2) the confounder-outcome relationships, and (3) the pollutant-confounder relationships. Second, we must assume that no additional confounders would be introduced in the no-regulation scenario.

3.2 Matching Estimator of the TEA

In this section, we introduce a causal inference matching procedure to estimate $\mathbb{E}(Y|\boldsymbol{X}_i, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2i}})$ (and thereby the TEA). In Section 4 of the Supplementary Materials, we show that the estimator is consistent, and we describe a bootstrapping approach that can be used to compute uncertainties. The idea behind our matching approach is simple: we will find all units with observed pollutant levels approximately equal to $\boldsymbol{t}_{2i}$ and confounder levels approximately equal to $\boldsymbol{X}_i$, and we will take the average observed outcome value across these units (plus a bias correction term) as an estimate of $\mathbb{E}(Y|\boldsymbol{X}_i, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2i}})$.

Let $\boldsymbol{\omega}$ be a Q-length vector of pre-specified constants and ν be a pre-specified scalar. For a column vector $\boldsymbol{b}$, $|\boldsymbol{b}|$ denotes the component-wise absolute value and $||\boldsymbol{b}|| = (\boldsymbol{b}'\boldsymbol{A}\boldsymbol{b})^{1/2}$, with $\boldsymbol{A}$ a positive semidefinite matrix. In practice, $\boldsymbol{A}$ will be the sample covariance matrix so that $||\boldsymbol{b}_1 - \boldsymbol{b}_2||$ is the Mahalanobis distance. We let $\varphi(i)$ denote the set of indices of the units matched to unit i. Then

$$\varphi(i) = \{j \in 1, ..., N : |\boldsymbol{t}_{2i} - \boldsymbol{t}_{1j}| \prec \boldsymbol{\omega} \ \wedge \ ||\boldsymbol{X}_i - \boldsymbol{X}_j|| < \nu\}$$

In the first condition here, we are carrying out exact matching within some tolerances $\boldsymbol{\omega}$ on the pollution variables, so that the units matched to unit i have observed pollution levels ($\boldsymbol{t}_{1j}$) almost equal to the counterfactual pollution levels for unit i ($\boldsymbol{t}_{2i}$). This ensures that the matched units have observed pollution values within a small hyperrectangle around $\boldsymbol{t}_{2i}$, i.e. $\boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2i}}$. The second condition carries out Mahalanobis distance matching within some tolerance ν on the confounders, so that all matched units have confounder values approximately equal to the confounder values for unit i. We use separate procedures for matching on the pollution and the confounder values so that we can exercise more

direct control over the closeness of the matches on pollution values. Literature on choosing calipers in standard matching procedures (Lunt, 2013) and identifying regions of common support in causal inference (King and Zeng, 2006) may provide insight into how to specify $\boldsymbol{\omega}$ and ν. All units that meet the above conditions are used as matches for unit i, thus each i is allowed to have a different number of matches, M_i. Matching is performed with replacement across the i. Then, we estimate $\mathbb{E}(Y|\boldsymbol{X}_i, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2i}})$ as

$$\hat{\mathbb{E}}(Y|\boldsymbol{X}_i, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2i}}) = \frac{P_i}{M_i} \sum_{k \in \varphi(i)} Y_k^* \tag{2}$$

We also test varieties of this matching estimator with bias corrections following Abadie and Imbens (2011), because bias-corrected matching estimators can provide increased robustness. Let $\hat{\mu}(\boldsymbol{t}_{2i}, \boldsymbol{X}_i)$ be a regression-predicted value of Y conditional on $\boldsymbol{t}_{2i}$ and $\boldsymbol{X}_i$. Then the bias-corrected estimator has the form

$$\hat{\mathbb{E}}(Y|\boldsymbol{X}_i, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2i}}) = \frac{1}{M_i} \sum_{k \in \varphi(i)} (Y_k^* P_i + \hat{\mu}(\boldsymbol{t}_{2i}, \boldsymbol{X}_i) - \hat{\mu}(\boldsymbol{t}_{2i}, \boldsymbol{X}_k))$$

While asymptotically matches will be available for all units (as a result of the overlap assumption, *A3*), in practice there will likely be some units for which no suitable matches can be found in the data. Let $s = 1, ..., S$ index the units for which 1 or more matches was found. Instead of τ, in finite sample settings we will generally estimate

$$\tau^* = \sum_{s=1}^{S} \mathbb{E}(Y|\boldsymbol{X}_s, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2s}}) - Y_s(\boldsymbol{T} = \boldsymbol{t}_{1s})$$

by plugging in $\hat{\mathbb{E}}(Y|\boldsymbol{X}_s, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2s}})$ for $\mathbb{E}(Y|\boldsymbol{X}_s, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2s}})$. This discarding of unmatched

units is often called "trimming" and the remaining S units called the trimmed sample. Trimming is done to avoid using extrapolation to estimate counterfactual outcomes in areas of the treatment/confounder space that are far from the observed data. Such extrapolation can produce results that are highly biased or model-dependent. The impacts of trimming on the interpretation of the TEA estimate are discussed at length in subsequent sections and in Section 5 of the Supplementary Materials.

3.3 BART for TEA Estimation

In this section, we propose an alternative approach to estimation of the TEA relying on machine learning. Machine learning procedures are typically applied in causal inference as a mechanism for imputation of missing counterfactual outcomes. BART (Chipman et al., 2010) is a Bayesian tree ensemble method. Several recent papers have shown its promising performance in causal inference contexts (Hill, 2011; Hahn et al., 2020).

The general form of the BART model for a continuous outcome Y and a vector of predictors $\boldsymbol{X}$ is $Y = \sum_{j=1}^{J} g(\boldsymbol{X}; \mathcal{T}_j, \mathcal{M}_j) + \epsilon$, where $j = 1, ..., J$ indexes the trees in the ensemble, g is a function that sorts each unit into one of a set of m_j terminal nodes, associated with mean parameters $\mathcal{M}_j = \{\mu_1, ..., \mu_{m_j}\}$, based on a set of decision rules, $\mathcal{T}_j$. $\epsilon \sim N(0, \sigma^2)$ is a random error term. BART is fit using a Bayesian backfitting algorithm. For estimation of average treatment effects (ATE) in a binary treatment setting, BART is fit to all the observed data and used to estimate the potential outcome under treatment and under control for each unit (Hill, 2011). The average difference in estimated potential outcomes is computed across the sample to obtain an ATE estimate.

We invoke BART similarly to estimate the TEA, inserting the rates as the outcome. For clarity in this section, we use the notation Y_i^* as the observed outcome rate under

the factual pollution levels and $\bar{Y}_i^*$ as the missing counterfactual outcome rate under the counterfactual pollution levels. Boldface versions denote the vector of all outcomes. We fit the following BART model to our data: $Y_i^* = \sum_{j=1}^{J} g(\boldsymbol{t}_{1i}, \boldsymbol{X}_i; \mathcal{T}_j, \mathcal{M}_j) + \epsilon_i$. We collect H posterior samples of the parameters from this BART model, referred to collectively as θ. For a given posterior sample h and for each unit i, we collect a sample from the posterior predictive distribution of $\bar{Y}_i^*$, $p(\bar{Y}_i^*|\mathbf{Y}^*) = \int p(\bar{Y}_i^*|\mathbf{Y}^*, \theta)p(\theta|\mathbf{Y}^*)d\theta$. Denote this posterior predictive sample $\bar{Y}_i^{*(h)}$. We use these to construct a posterior predictive sample of τ as $\tau^{(h)} = \sum_{i=1}^{N}(\bar{Y}_i^{*(h)} - Y_i^*)P_i$. We then estimate $\hat{\tau} = \frac{1}{H}\sum_{h=1}^{H}\tau^{(h)}$, i.e. the posterior mean. To relate this back to the notation of the previous section, note that this is equivalent to estimating $\hat{\mathbb{E}}(Y|\boldsymbol{X}_i, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2i}}) = \frac{1}{H}\sum_{h=1}^{H}\bar{Y}_i^{*(h)}P_i$ for all i. A 95% credible interval is formed with the 2.5% and 97.5% percentiles of the $\tau^{(h)}$.

Recall that, with the matching estimator, units for which no matches can be found within the pre-specified tolerances are trimmed to avoid extrapolation, so that we estimate τ^* rather than τ. BART does not automatically identify units for which extrapolation is necessary to estimate counterfactual outcomes. Hill and Su (2013) proposed a two-step method to identify units for trimming with BART which could be adapted for use here. In our simulations and analyses, to ensure comparability of the results from BART and matching, we fit the BART model to the entire dataset but, in estimating the TEA, we omit the same set of units that are trimmed with matching.

4 Simulations

In this section, we generate synthetic data mimicking the structure of our real data and test the following methods for estimating τ^*: our matching procedure with various types of bias

correction, our BART procedure, and a simple Poisson regression. In these simulations, we let $N = 4,900$ and $Q = 2$, with one pollutant simulated to mimic $PM_{2.5}$ and the other O_3. The observed pollutants are generated as spatial processes on a 70×70 spatial grid using the `gstat` package in R (Gräler et al., 2016). The vectors of $PM_{2.5}$ and O_3 values ($\mathbf{PM}$ and $\mathbf{O}$) are generated from $\mathbf{PM} \sim MVN(12.18, \boldsymbol{\Sigma}_1)$ and $\mathbf{O} \sim MVN(50, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_1$ are exponential spatial covariance matrices. Then, letting $\boldsymbol{t}_{1i} = [PM_i \;\; O_i]$, we generate the counterfactual pollution levels as $\boldsymbol{t}_{2i} = \boldsymbol{t}_{1i} + \boldsymbol{z}_i$, $\boldsymbol{z}_i = [z_{1i} \;\; z_{2i}]$, $z_{1i} \sim Unif(0, 6.64)$ and $z_{2i} \sim Unif(0, 25.87)$. The result is that the counterfactual pollutant levels are always larger than the observed pollutant levels, but the magnitudes of the differences vary across units and at appropriate scales for each pollutant.

We use the exposure values to generate five confounders, $X_{1i}, ..., X_{5i}$. We let $X_{hi} = \boldsymbol{t}'_{1i}\boldsymbol{\alpha}_h + \epsilon_{hi}$, where $\epsilon_{hi} \sim N(0, \sigma_h^2)$. In addition to the exposures and confounders, we also generate four random variables used as "unobserved" predictors of Y, i.e. we use them to generate Y but do not include them in the estimation procedures (they are not confounders because they are not related to exposure). In our real data analysis, there are likely many unobserved predictors of the health outcomes considered (but hopefully no unobserved confounders), thus it is important to evaluate our methods under such conditions.

In each simulation, we generate $Y_i \sim Poisson(\lambda_i)$, with λ_i a function of the observed exposures, confounders, and predictors. We also generate an expected counterfactual outcome for each unit, λ_i^{cf} using the same functional form but substituting the counterfactual exposure values for the observed exposure values. We consider three different functional forms for λ_i, and we refer to the different structures as S-1, S-2, and S-3. S-1 uses the most complex form to construct Y, involving strong non-linearities and interactions both within and across the exposures and confounders. S-2 is slightly less complex, with exposure-

confounder interactions excluded. In S-3, all relationships are linear. See the model forms and parameter values used in each simulation in Section 8 of the Supplementary Materials.

With these data, we first perform matching to determine which units have suitable matches for their $\boldsymbol{t}_{2i}$ in the data and will therefore contribute to the estimation of τ^*. Within each of the three simulation scenarios, we set $\nu = 1.94$, which is approximately the 10^{th} percentile of the Mahalanobis distances between all the units in the data, and consider three different specifications of $\boldsymbol{\omega}$, the tolerances for the pollutant matches. We use 15%, 20%, and 25% of the standard deviations of the counterfactual pollutant distributions, resulting in $\boldsymbol{\omega} = [0.49 \;\; 1.92]$, $\boldsymbol{\omega} = [0.65 \;\; 2.56]$, and $\boldsymbol{\omega} = [0.82 \;\; 3.20]$. Following matching, we estimate three variations of the matching estimator. The first is the simple matching estimator given in equation 2 (Match 1). The second is this matching estimator plus a bias correction from a Poisson generalized additive model (GAM) with smooth terms on each predictor, i.e. $\log(\lambda) = \sum_{j=1}^{5} (s_j(X_j)) + s_6(t_{11}) + s_7(t_{12})$ (Match 2). The third is the matching estimator with a bias correction from a Poisson regression in which the forms of the exposures and confounders are correctly specified (Match 3). We also apply BART as described in Section 3.3, fitting the BART model to all the data but estimating the TEA with only the units retained after matching. We also compare these methods to a simple Poisson regression. We fit a Poisson regression with all exposures and confounders included as linear terms (PR 1) and a Poisson regression model in which the forms of the exposures and confounders are correctly specified (PR 2). We compare each estimate to the true TEA in the trimmed sample, $\tau^* = \sum_{s=1}^{S} \lambda_s^{cf} - \lambda_s$.

Table 1 contains the results of 200 simulations. It shows the proportion of units retained after trimming, the ratio of the true trimmed sample TEA to the true whole sample TEA, and the percent bias and 95% confidence/credible interval (CI) coverage for each method

Table 1: Percent absolute bias (95% CI coverage) of the following methods in estimation of τ^* in 200 simulations: matching with no bias correction (Match 1), matching with a GAM bias correction (Match 2), matching with correctly specified bias correction (Match 3), BART, Poisson regression with linear terms (PR 1), correctly specified Poisson regression (PR 2). $\boldsymbol{\omega}$ refers to matching tolerances on pollutants, i.e. 0.15 means that matches are restricted to be within 0.15 standard deviations for each pollutant. S/N is the proportion of the sample retained after trimming, and τ^*/τ is the ratio of the true trimmed sample TEA to the true whole sample TEA.

	$\boldsymbol{\omega}$	S/N	τ^*/τ	Match 1	Match 2	Match 3	BART	PR 1	PR 2
S-1 $\tau = 195011$	0.15	0.51	0.07	0.42 (0.00)	0.01 (0.98)	0.01 (0.93)	0.05 (0.62)	0.22 (0.00)	0.01 (0.88)
	0.20	0.58	0.09	0.42 (0.00)	0.00 (0.98)	0.05 (0.46)	0.06 (0.57)	0.21 (0.00)	0.02 (0.86)
	0.25	0.63	0.11	0.38 (0.00)	0.04 (0.74)	0.09 (0.08)	0.06 (0.56)	0.20 (0.00)	0.02 (0.81)
S-2 $\tau = 186852$	0.15	0.51	0.20	0.04 (1.00)	0.02 (0.81)	0.02 (0.8)	0.00 (0.90)	0.05 (0.12)	0.01 (0.87)
	0.20	0.58	0.25	0.03 (1.00)	0.03 (0.42)	0.03 (0.39)	0.00 (0.90)	0.06 (0.04)	0.02 (0.77)
	0.25	0.63	0.29	0.01 (0.94)	0.06 (0.08)	0.06 (0.06)	0.01 (0.92)	0.07 (0.00)	0.02 (0.80)
S-3 $\tau = 132818$	0.15	0.51	0.41	0.29 (0.00)	0.04 (0.98)	0.03 (0.98)	0.03 (0.99)	0.04 (0.53)	-
	0.20	0.58	0.47	0.30 (0.00)	0.04 (1.00)	0.04 (1.00)	0.05 (0.97)	0.06 (0.34)	-
	0.25	0.63	0.53	0.29 (0.00)	0.04 (0.98)	0.03 (0.99)	0.07 (0.89)	0.07 (0.18)	-

in each simulation. Across all simulations, substantial portions of the sample are being trimmed (37%-49%). These trimmed units account for a disproportionately high amount of the TEA, as τ^*/τ is generally much smaller than S/N. This reflects the scenario we would expect in our real data, because the units with the highest counterfactual pollution levels are likely to have the largest effect sizes, yet these units are unlikely to be matched since we may observe few or no pollution exposures as high as their counterfactuals.

In practice we generally do not know the correct model form, therefore the results of the oracle models in Match 3 and PR 2 have little relevance and we will focus on a comparison of the other methods. Among the remaining methods, we generally see that as $\boldsymbol{\omega}$ increases, the bias of the TEA estimates increase. This is consistent with expectations, as a larger $\boldsymbol{\omega}$ allows matches with more distant exposure values, which can lead to inappropriate extrapolation.

Match 1 and PR 1 each give highly biased results under certain data generating mechanisms and, in all but one case, have bias greater than or equal to that of BART and Match 2.

Both BART and Match 2 have strong and consistent performance across all data generating mechanisms, with bias $\leq 7\%$ in all simulations. Match 2 generally has lower bias in S-1 and S-3, while BART has lower bis in S-2. The trends in coverage of Match 2 are consistent with the bias, i.e., when bias is low, $\geq 95\%$ coverage tends to be achieved and, as the bias increases, the coverage experiences a corresponding decrease. With the more complex data generating mechanisms (S-1 and S-2), BART tends to give unsatisfactory coverage, even when bias is low. This anti-conservative property of BART predictive uncertainties in settings with complex data generating processes is consistent with previous findings (Wendling et al., 2018; Hahn et al., 2020), and appears to be a common limitation of BART which should be considered when making inference.

These results demonstrate the trade-offs that must be considered when choosing the matching tolerances, i.e., when choosing which units to trim. Because the TEA is a sum, the removal of one unit when estimating the TEA is likely to have a larger impact than it would when estimating an average. The stricter the tolerances, the more units are dropped from the analysis and, generally, the more distant τ^* will grow from τ. Therefore, the estimated TEA is less and less likely to reflect the population of interest. In our context where greater pollution exposure is unlikely to have protective effects on health, and where most of the counterfactual pollution levels are greater than the factual ones (i.e. $Y(\boldsymbol{T} = \boldsymbol{t}_{2i}) - Y(\boldsymbol{T} = \boldsymbol{t}_{1i}) > 0$), we anticipate that stricter tolerances and the trimming of more units will lead to underestimation of the TEA. However, stricter tolerances reduce the potential for extrapolation and generally give estimates of τ^* with lower bias. The selection of these tolerances should be considered carefully in the context of each data application.

5 Application

We return now to the real data described in Section 2. Because the purpose of the CAAA was to reduce air pollution, we would expect the counterfactual (no-CAAA) pollution exposures to be higher than factual (with-CAAA) pollution exposures in most zipcodes; however, the CAAA resulted in a complex set of regulations that could have led to decreases in some areas and increases in others. In our data, in 30% of zipcodes the factual $PM_{2.5}$ and/or O_3 exposure estimate is larger than the corresponding counterfactual estimate, indicating that the CAAA increased exposure. See Section 8 of the Supplementary Materials for a map of these zipcodes. This could reflect real increases in pollution exposures due to the CAAA or it could be partially due to differences in the pollution exposure modeling used to produce the factual and counterfactual estimates. In our primary analysis (PA), to be conservative we include these zipcodes, and we also perform a sensitivity analysis (SA) in which they are removed. In the SA, we allow all zipcodes to serve as potential matches but the TEA's sum is only taken across zipcodes with both factual pollutant estimates smaller than the corresponding counterfactual.

Within each of the PA and SA, we conduct separate analyses to estimate the health impacts in years 2000 and 2001 due to CAAA-attributable changes in pollution exposures in the year 2000. Prior to analysis, we remove any zipcodes with missing data or with zero Medicare population. For the year 2000, the sample size following these removals is $N = 28,724$ zipcodes (SA: $N = 20,027$). All descriptive statistics provided are from the year 2000 data. Figures for 2001 deviate little, if at all, due to year-specific missingness and Medicare person-time. For each year's data we apply BART, matching with a GAM bias correction, and a Poisson regression with linear terms to estimate the TEA for mortality, cardiovascular disease (CVD) hospitalizations, and dementia hospitalizations.

We first apply matching with a GAM bias correction to the data, collecting 50 bootstrap replicates and setting tolerances $\boldsymbol{\omega} = [0.50 \;\; 0.77]$ and $\nu = 2.78$. The $\boldsymbol{\omega}$ values are 0.1 standard deviations of each counterfactual pollutant distribution, and ν is approximately the 10^{th} percentile of the Mahalanobis distances between all the confounders in the data. These tolerances lead to trimming of 11,425 zipcodes (SA: 10,196), leaving $S = 17,299$ zipcodes (SA: $S = 9,831$) for estimation of the TEA. The portion of the full dataset retained to estimate the TEA, $S/N = 0.60$ (SA: $S/N = 0.49$), is similar to that in the simulations. As in Section 3, we fit the BART and Poisson regression to the entire sample but only include the units retained after trimming for TEA estimation.

The total Medicare population in the zipcodes retained after trimming is 14,880,606 (SA: 7,820,770). Table 2 shows the means and standard deviations of the health outcome rates, exposures and confounders in the full dataset, among the discarded/trimmed zipcodes, and in the retained/untrimmed zipcodes used for estimation for the PA. In Section 8 of the Supplementary Materials, we provide a map of the discarded and retained zipcodes after trimming for the PA, and an analogue of Table 2 for the SA. The discarded zipcodes are primarily in the mid-Atlantic region, where Figure 1 shows some of the highest counterfactual pollutant levels, and in major northeast and west coast population centers. Notably, the discarded zipcodes have larger average population size and higher average counterfactual pollutant exposures than the full dataset. This is likely due to the fact that we have sparse observed data at very high exposure levels and therefore it is difficult to find matches for zipcodes with very high counterfactual exposures (which also tend to be the more populous zipcodes). We would also expect that many of the largest health benefits of the CAAA may have come in zipcodes with large populations and whose no-CAAA exposures would have been very high, thus our results are likely to be underestimates of

the entire health benefits of the CAAA. See Section 5 of the Supplementary Materials for a more detail on the interpretation of the TEA estimate in this context. Additionally, because the sample retained after trimming under-represents some regions and population subgroups, we discourage the use of this approach to make comparisons across such groups, e.g., for the purpose of examining environmental justice issues.

Table 2: Average (and standard deviation) of Medicare population size, Medicare health outcome rates, pollutant exposures and confounders in the full dataset, only the discarded/trimmed zipcodes, and the retained/untrimmed zipcodes used for estimation.

	Full Dataset	Discarded Units	Retained Units
Medicare population size	1102.78 (1537.59)	1470.08 (1815.02)	860.2 (1265.74)
FFS person-months	10791.81 (14548.64)	13723.07 (16799.63)	8855.88 (12475.97)
Mortality (rate per 1,000)	52.19 (21.33)	52.66 (22.19)	51.88 (20.75)
Dementia (rate per 1,000)	1.58 (1.28)	1.7 (1.46)	1.5 (1.15)
CVD (rate per 1,000)	6.77 (2.71)	6.8 (2.8)	6.75 (2.65)
Factual $PM_{2.5}$ ($\mu g/m^3$)	10.48 (3.79)	11.47 (3.61)	9.83 (3.76)
Factual O_3 (ppb)	47.72 (6.58)	47.08 (6.87)	48.15 (6.34)
Counterfactual $PM_{2.5}$ ($\mu g/m^3$)	14.12 (5.01)	17.32 (4.7)	12.01 (3.99)
Counterfactual O_3 (ppb)	53.57 (7.69)	57.85 (7.56)	50.74 (6.34)
poverty (proportion)	0.11 (0.1)	0.11 (0.11)	0.11 (0.08)
popdensity (per mi^2)	1155.93 (4396.92)	2319.84 (6676.24)	387.23 (1086.17)
housevalue (USD)	104021.52 (83107.11)	133984.69 (111102.05)	84232.56 (48290.31)
black (proportion)	0.08 (0.16)	0.09 (0.18)	0.06 (0.14)
income (USD)	40193.35 (15989.31)	45151 (19837.6)	36919.11 (11733.04)
ownhome (proportion)	0.75 (0.15)	0.72 (0.19)	0.78 (0.11)
hispanic (proportion)	0.06 (0.12)	0.08 (0.16)	0.05 (0.1)
education (proportion)	0.38 (0.18)	0.37 (0.19)	0.39 (0.17)
northeast (proportion)	0.18 (0.39)	0.37 (0.48)	0.06 (0.23)
midwest (proportion)	0.32 (0.47)	0.26 (0.44)	0.36 (0.48)
south (proportion)	0.34 (0.47)	0.23 (0.42)	0.42 (0.49)
west (proportion)	0.16 (0.36)	0.14 (0.35)	0.16 (0.37)

The results of both the PA and SA appear in Figure 2. The point estimates in the PA

and SA are similar, but the 95% CIs are substantially wider in the PA. This suggests that the zipcodes for which one or both of the factual pollutant exposures are larger than the counterfactual are adding noise into the results. Matching and BART find limited, although somewhat inconsistent, evidence of effects on mortality. Only the BART SA for 2000 detects an effect, estimating approximately 8,000 mortalities prevented in the retained zipcodes, with the lower CI limit just exceeding zero. For CVD and dementia hospitalizations, all the methods yield large positive point estimates, i.e., large estimates of the number of events avoided in the specified year due to CAAA-attributable pollution changes in 2000. The estimates from BART and matching suggest that approximately 15,000 dementia hospitalizations and 15,000-25,000 CVD hospitalizations were avoided in each year. Nearly all of the results for CVD and dementia are "statistically significant", i.e. the 95% CIs do not overlap zero, providing strong evidence of an effect. The Poisson regression gives positive and statistically significant TEA estimates for all outcomes, with the significance attributable to extremely narrow CIs. We ran additional sensitivity analyses to evaluate robustness to potential error in the pollution exposures and to the choice of confounder matching parameter ν. Results are provided in Section 6 of the Supplementary Materials. In summary, these analyses suggest that our findings are robust to these factors.

Air pollution literature accumulated over the past 20 years has provided strong evidence to support a causal relationship between long-term $PM_{2.5}$ exposure and mortality (US EPA, 2019), including studies using Medicare data (Di et al., 2017). One major difference in our study compared to previous cohort studies examining long-term $PM_{2.5}$ exposure and mortality is that we are using zipcode aggregated data while they use individual level data. The use of individual level data allows for adjustments for individual level features, while the use of aggregated data does not. Moreover, due to the potential for ecologic bias,

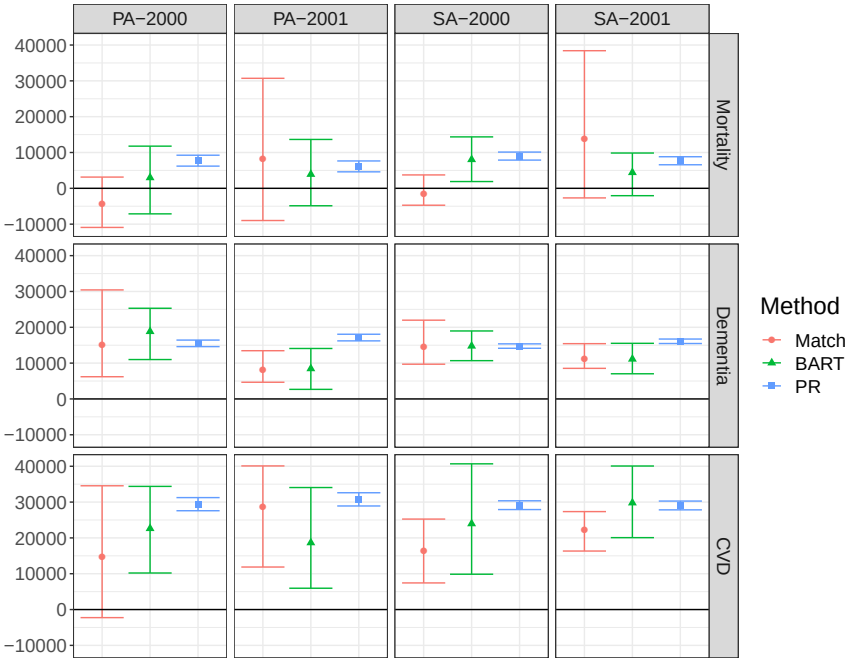

Figure 2: Primary analysis (PA) and sensitivity analysis (SA) estimates and 95% CIs for the TEA in the Medicare population in 2000 and 2001 due CAAA-attributable changes in $PM_{2.5}$ and O_3 in the year 2000. Estimation is performed using matching with a GAM bias correction, BART, and Poisson regression (PR).

inference from the analysis of aggregate spatial data is inextricably linked to the spatial units of analysis and cannot be transported to the individual level. Therefore, results from individual level data are often preferred. Given the complex nature of mortality, our inability to adjust for individual level factors may make mortality rates appear too noisy to detect pollution effects. Our results are consistent with the those of Henneman et al. (2019) who study zipcode level coal emissions exposures and find limited effects on mortality but strong evidence of harmful effects on CVD hospitalizations in the Medicare population.

To demonstrate how the approach introduced in this paper can be used in conjunction

with traditional health impact assessments, we provide the EPA's Section 812 Analysis estimates for the number of mortalities and CVD hospitalizations prevented due to the CAAA in Section 7 of the Supplementary Materials. We also provide context to clarify how their estimates can be compared to ours. In short, our approach yields results entirely supported by real, population-level health outcome data, but that are likely conservative due to trimming. Our methods also account for any synergistic effects of the pollutants. The traditional approach used in the Section 812 analysis relies on 1) health effect estimates from cohort studies which allow for extensive confounding adjustment but may not fully reflect the population of interest; and 2) a stronger set of modeling assumptions which, while unverifiable, allows for the number of events avoided to be estimated in all locations, even areas with no support in the real data. The results of both types of analysis provide useful insights about the health impacts of large air pollution regulations.

The strong causal identifying assumptions in Section 3 must be satisfied to obtain causal results. The unconfoundedness assumption may be violated, because we have not directly adjusted for potential behavioral confounders, e.g., smoking. However, these features may be correlated with features we do adjust for, e.g., education or region. We also make the strong assumption of no interference. It is unclear how a violation of this assumption would impact our results. However, this assumption has been made in most causal inference analyses of air pollution to date (Papadogeorgou and Dominici, 2018; Wu et al., 2018).

6 Discussion

In this paper, we have introduced a causal inference approach for evaluating the health impacts of an air pollution regulation. We developed an estimand called the TEA and pro-

posed a matching and a machine learning method for estimation. Both methods showed promising performance in simulations. We implemented these methods to estimate the TEA for mortality, dementia hospitalizations, and CVD hospitalizations in the Medicare population due to CAAA-attributable pollution changes in the year 2000. We found compelling evidence that CAAA-attributable pollution changes prevented large numbers of CVD and dementia hospitalizations. Because more than one third of the zipcodes were trimmed from our analysis to avoid extrapolation, and because the trimmed zipcodes tend to have larger populations and larger improvements in air quality, we expect that the true number of health events avoided may be considerably larger than our estimates.

While our causal inference approach improves on the traditional approach to regulation evaluation in many ways, there are trade-offs to be considered. In order to avoid extrapolation and strong model-dependence, our methods discard units whose counterfactual pollution/confounder values are far outside the observed pollution/confounder space. This often leads to discarding of units where the highest impacts would be expected. Our estimates tend to have small bias in estimating the effects in the trimmed sample; however the effects in the trimmed sample may be much lower than in the entire original sample. The traditional approach to regulatory evaluation retains all units for estimation, and uses parametric models to extrapolate counterfactual outcomes for units with pollutant/confounder values outside areas of support in the observed data. This extrapolation could produce biased estimates, but it is not obvious what the direction of that bias would be. It may be useful to apply and compare both approaches in future regulation evaluations.

The limitations of our methods present opportunities for future methodological advancements. In particular, approaches for causal inference with interference could be integrated from recent work (Barkley et al., 2017; Papadogeorgou et al., 2019). To our knowledge, no

methods yet exist that can formally account for spatially dependent data in the context of matching or tree ensemble models, and such extensions could improve model stability and uncertainty estimates in this context. We also note that, because our outcome is rate data, BART's normality assumption may be violated. A BART for count data has been proposed (Murray, 2017), but at the time of writing this manuscript, code was not available. Moreover, future work could extend our approach to accommodate individual level data instead of aggregated data. Finally, our approach relies on pollution data from different sources, i.e., factual pollution estimates from hybrid models and counterfactual estimates from atmospheric chemistry models. Although we justify this choice, our results may suffer from incompatibility in these data sources, which could be improved upon in future work.

Finally, improvements in pollution exposure data more broadly are essential to increase the reliability of our approach. Statisticians and engineers should work together to produce nationwide factual and counterfactual pollution exposure estimates that are increasingly spatially granular and are more tailored for this type of analysis. These estimates should be accompanied by uncertainties, which could be taken into account in downstream analyses.

Disclaimer: This manuscript has been reviewed by the U.S. Environmental Protection Agency and approved for publication. The views expressed in this manuscript are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

SUPPLEMENTARY MATERIAL

The Supplementary Material contains additional information about traditional approaches to regulation health impact evaluations, our matching estimator, the simulations, the CAAA analysis, and the Section 812 analysis results. (PDF file)

References

Abadie, A. and G. Imbens (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics 29*(1), 1–11.

Abu Awad, Y., Q. Di, Y. Wang, C. Choirat, B. Coull, A. Zanobetti, and J. Schwartz (2019). Change in $PM_{2.5}$ exposure and mortality among Medicare recipients. *In press, Environmental Epidemiology*.

Barkley, B. G., M. G. Hudgens, J. D. Clemens, M. Ali, and M. E. Emch (2017). Causal inference from observational studies with clustered interference. *arXiv preprint arXiv:1711.04834*.

Brook, R. D., S. Rajagopalan, C. A. Pope III, J. R. Brook, A. Bhatnagar, A. V. Diez-Roux, F. Holguin, Y. Hong, R. V. Luepker, M. A. Mittleman, et al. (2010). Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. *Circulation 121*(21), 2331–2378.

Chipman, H. A., E. I. George, R. E. McCulloch, et al. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics 4*(1), 266–298.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Di, Q., S. Rowland, P. Koutrakis, and J. Schwartz (2017). A hybrid model for spatially and temporally resolved ozone exposures in the continental united states. *Journal of the Air & Waste Management Association 67*(1), 39–52.

Di, Q., Y. Wang, A. Zanobetti, Y. Wang, P. Koutrakis, C. Choirat, F. Dominici, and J. D.

Schwartz (2017). Air pollution and mortality in the Medicare population. *New England Journal of Medicine 376*(26), 2513–2522.

Gräler, B., E. Pebesma, and G. Heuvelink (2016). Spatio-temporal interpolation using gstat. *The R Journal 8*, 204–218.

Hahn, P. R., J. S. Murray, C. M. Carvalho, et al. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*.

Henneman, L. R., C. Choirat, and C. M. Zigler (2019). Accountability assessment of health improvements in the United States associated with reduced coal emissions between 2005 and 2012. *Epidemiology 30*(4), 477–485.

Henneman, L. R., C. Liu, H. Chang, J. Mulholland, P. Tolbert, and A. Russell (2019). Air quality accountability: Developing long-term daily time series of pollutant changes and uncertainties in Atlanta, Georgia resulting from the 1990 Clean Air Act Amendments. *Environment International 123*, 522–534.

Hill, J. and Y.-S. Su (2013). Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, 1386–1420.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics 20*(1), 217–240.

Ho, D. E., K. Imai, G. King, and E. A. Stuart (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis 15*(3), 199–236.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika 87*(3), 706–710.

King, G. and L. Zeng (2006). The dangers of extreme counterfactuals. *Political Analysis 14*(2), 131–159.

Kreif, N., R. Grieve, I. Díaz, and D. Harrison (2015). Evaluation of the effect of a continuous treatment: A machine learning approach with an application to treatment for traumatic brain injury. *Health Economics 24*(9), 1213–1228.

Lunt, M. (2013). Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *American Journal of Epidemiology 179*(2), 226–235.

Murray, J. S. (2017). Log-linear Bayesian additive regression trees for categorical and count responses. *arXiv preprint arXiv:1701.01503*.

Nethery, R. C. and F. Dominici (2019). Estimating pollution-attributable mortality at the regional and global scales: challenges in uncertainty estimation and causal inference. *European Heart Journal*.

Papadogeorgou, G. and F. Dominici (2018). A causal exposure response function with local adjustment for confounding: A study of the health effects of long-term exposure to low levels of fine particulate matter. *arXiv preprint arXiv:1806.00928*.

Papadogeorgou, G., F. Mealli, and C. M. Zigler (2019). Causal inference with interfering units for cluster and population level treatment allocation programs. *In press, Biometrics*.

Pope III, C. A., R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, K. Ito, and G. D. Thurston (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama 287*(9), 1132–1141.

Power, M. C., S. D. Adar, J. D. Yanosky, and J. Weuve (2016). Exposure to air pollution as a potential contributor to cognitive function, cognitive decline, brain imaging, and dementia: a systematic review of epidemiologic research. *Neurotoxicology 56*, 235–253.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association 75*(371), 591–593.

Sacks, J. D., J. M. Lloyd, Y. Zhu, J. Anderton, C. J. Jang, B. Hubbell, and N. Fann (2018). The Environmental Benefits Mapping and Analysis Program–Community Edition (BenMAP–CE): A tool to estimate the health and economic benefits of reducing air pollution. *Environmental Modelling & Software 104*, 118–129.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics 25*(1), 1–21.

US EPA (2011). Benefits and Costs of the Clean Air Act 1990-2020, the Second Prospective Study. Accessed Online: 2019-05-08.

US EPA (2019). Integrated Science Assessment (ISA) For Particulate Matter (Final Report, 2019). U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-19/188. Accessed Online: 2020-04-16.

van Donkelaar, A., R. V. Martin, C. Li, and R. T. Burnett (2019). Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology*.

van Erp, A. M., R. O'Keefe, A. J. Cohen, and J. Warren (2008). Evaluating the effectiveness of air quality interventions. *Journal of Toxicology and Environmental Health, Part A 71*(9-10), 583–587.

Wendling, T., K. Jung, A. Callahan, A. Schuler, N. Shah, and B. Gallego (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine 37*(23), 3309–3324.

WHO (2019). AirQ+: software tool for health risk assessment of air pollution. Accessed Online: 2019-08-17.

Wu, X., F. Mealli, M. Kioumourtzoglou, F. Dominici, and D. Braun (2018). Matching on generalized propensity scores with continuous exposures. *arXiv preprint arXiv:1812.06575*.

Zigler, C. M., F. Dominici, and Y. Wang (2012). Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics 13*(2), 289–302.

Zigler, C. M. and G. Papadogeorgou (2018). Bipartite causal inference with interference. *arXiv preprint arXiv:1807.08660*.